Chemometrics and Intelligent Laboratory Systems 147 (2015) 176–184

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Improvement of iterative optimization technology (for process analytical technology calibration-free/minimum approach) with dimensionality reduction and wavelength selection of spectra

Hiromasa Kaneko [a], Koji Muteki [b], Kimito Funatsu [a,*]

[a] Department of Chemical System Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[b] Pfizer Global Supply, Global Technology Services, Eastern Point Rd, Groton, CT 06340, United States

ARTICLE INFO

Article history:
Received 30 May 2015
Received in revised form 18 August 2015
Accepted 24 August 2015
Available online 30 August 2015

Keywords:
Process analytical technology
Iterative optimization technology
Dimensionality reduction
Wavelength selection

ABSTRACT

Process analytical technology plays an important role in the pharmaceutical industry. Calibration-free/minimum approach, iterative optimization technology (IOT), was previously proposed to predict the composition of a mixture while maintaining a similar prediction ability to calibration models such as a partial least squares. However, for the mixture case which includes similar structured materials, it would be essentially difficult to provide good prediction on mixture component ratio. This study presents a method which can improve the prediction ability of IOT through reducing dimensionality of spectra with optimal selection of wavelength. It involves using a latent variable model approach for dimensionality reduction of spectra in IOT and genetic algorithm-based wavelength selection for optimal wavelength-region selection. Through the analyses of numerical simulation data and real industrial data, it was confirmed that the proposed method achieved higher predictive accuracy compared to the traditional IOT.

1. Introduction

Process analytical technology (PAT) [1–4] plays an important role in the pharmaceutical industry. PAT is used extensively in process development, process understanding, and process control [5]. Often, quantitative measurements are desired/required and a calibrated model will have to be developed and implemented. A calibrated model could account for active pharmaceutical ingredient (API) lot-to-lot variability and scale-up operations and was applied to optimization of process conditions in pilot-scale plant and industrial-scale plant of the high shear wet granulation process [6]. García-Muñoz et al. proposed to use a calibrated model for feed-forward process control and for designing quality into process conditions and materials' properties (Quality by Design) [7]. The development, implementation and maintenance of these quantitative models are both resource and time intensive.

Muteki et al. [8] previously proposed calibration-free/minimum approach, iterative optimization technology (IOT), which is used to predict the composition of a mixture while maintaining a similar prediction ability to calibration models such as a partial least squares (PLS) [9] model. It involves using only pure component spectra and mixture component spectra (without a calibration data set). Linear and non-linear IOT have been successfully applied (by offline/online) to some practical pharmaceutical process (e.g. blending, feed frame, solvent mixture, reaction, etc).

However, a key (remaining) question was how to deal with a dependency (multi-collinear relationship) among pure component spectra during the optimization computation. For the mixture case which includes similar structured materials, it would be essentially difficult to provide good prediction on mixture component ratio, although the degree of collinearity of spectra can be changed by different measurement conditions of spectrometer. This problem has often been an important obstacle when applying IOT to mixture cases having more than several components. In addition, nonlinear relationships between the absorption of pure components and that of mixture components also deteriorate prediction ability of IOT. Although nonlinear IOT has been developed already [8], this requires to optimize a parameter using training samples and linear models would be more robust than nonlinear ones. Selection of only absorption bands with there exist linear relationships between the absorption of pure components and that of mixture components would be one of the solutions.

This study presents a method which can improve the prediction ability of IOT through reducing dimensionality of spectra with optimal selection of wavelength. It involves using a latent variable model (LVM) approach [10] such as principal component analysis (PCA) [11] and PLS for dimensionality reduction of spectra in IOT and genetic-algorithm-based wavelength selection (GAWLS) [12] for optimal wavelength-region selection. LVM is used as preprocessing of spectra and collinearity in pure component spectra would be handled. IOT

* Corresponding author. Tel.: +81 3 5841 7751; fax: +81 3 5841 7771.
E-mail address: funatsu@chemsys.t.u-tokyo.ac.jp (K. Funatsu).

http://dx.doi.org/10.1016/j.chemolab.2015.08.017
0169-7439/

estimation is applied to the calculation of a fitness value in GAWLS and a combination of wavelength-regions with which a predictive IOT model can be optimized would be obtained.

The effectiveness of the proposed methods is demonstrated using simulation data where pure component spectra have strong correlation and real industrial data of multicomponent mixtures. The proposed methods achieve higher predictive accuracy compared to the traditional IOT.

2. Method

2.1. LVM-IOT

Prediction ability of IOT decreases when there is collinearity between pure component spectra and collinearity between wavelengths (see Appendix A). To handle collinearity and guarantee independency of pure component spectra, our approach is dimensionality reduction and we apply latent variable modeling (LVM) as preprocess of IOT. The details of IOT are shown in Appendix A.

One of the most famous LVM methods is PCA, the details of which are described in Appendix B. PCA can reduce the dimension of **X** and extract uncorrelated PCs which are input variables for IOT.

PCA can be used in different ways. Fig. 1 shows images of two ways to employ PCA. First, principal components (PCs) are extracted from pure component spectra data matrix $\mathbf{X}_{\mathrm{pure}} \in R^{I \times N}$ (I represents the number of pure components and N represents the number of wavelengths). By using a loading matrix $\mathbf{P} \in R^{N \times q}$ (q represents the number of PCs), score vectors $\mathbf{T}_{\mathrm{mix}} \in R^{J \times q}$ (J represents the number of mixture component spectra) can be obtained from mixture component spectra data matrix $\mathbf{X}_{\mathrm{mix}} \in R^{J \times N}$ as follows:

$$\mathbf{T}_{\mathrm{mix}} = \mathbf{X}_{\mathrm{mix}}\mathbf{P}. \tag{1}$$

IOT can estimate mole fractions $\mathbf{R} \in R^{I \times J}$ using $\mathbf{T}_{\mathrm{pure}} \in R^{I \times q}$ instead of $\mathbf{X}_{\mathrm{pure}}$ and $\mathbf{T}_{\mathrm{mix}}$ instead of $\mathbf{X}_{\mathrm{mix}}$ in Eqs. (A.7) and (A.8). This method is named PCA-IOT1. PCA-IOT1 can handle correlation between wavelengths.

Second, PCs are extracted from pure component spectra data matrix $\mathbf{X}^{\mathrm{T}}_{\mathrm{pure}} \in R^{N \times I}$ after transposition (lower part in Fig. 1). IOT can estimate **R** for $\mathbf{T}_{\mathrm{pure}} \in R^{N \times q}$, $\mathbf{R}_{\mathrm{Tpure}} \in R^{q \times J}$, using $\mathbf{T}^{\mathrm{T}}_{\mathrm{pure}}$ instead of $\mathbf{X}_{\mathrm{pure}}$ and $\mathbf{X}_{\mathrm{mix}}$ in Eqs. (A.7) and (A.8). $\mathbf{R}_{\mathrm{Tpure}}$ is transformed to **R** using $\mathbf{P} \in R^{I \times q}$ as follows:

$$\mathbf{R} = \mathbf{P}\mathbf{R}_{\mathrm{Tpure}}. \tag{2}$$

This method is named PCA-IOT2. PCA-IOT2 can handle correlation between pure spectra.

For PCA-IOT1, PCA can be applied also to the data matrix combined with $\mathbf{X}_{\mathrm{pure}}$ and $\mathbf{X}_{\mathrm{mix}}$, and $\mathbf{T}_{\mathrm{pure}}$ and $\mathbf{T}_{\mathrm{mix}}$ are extracted simultaneously. This is named PCA-IOT3. PLS instead of PCA can be performed as LVM. PLS requires objective variables $\mathbf{Y} \in R^{J \times I}$. The mole fraction of each pure component is 1 for **Y**-variables, which is also used when Amigo et al. applied PLS to classical least squares [13]. PLS is applied as is the case with PCA-IOT1. This is named PLS-IOT.

2.2. GAWLS-IOT

Except an LVM approach, selection of the suitable absorption bands is another solution to guarantee the independency of pure component spectra. In addition, absorption bands having nonlinearity between pure component spectra and mixture component spectra, which come from molecular interactions such as hydrogen bonding, deteriorate prediction ability of IOT. Unnecessary wavelengths must be removed from pure component spectra and mixture component spectra.

Our second approach is wavelength selection or variable selection. Many variable selection methods such as uninformative variable elimination [14], searching combination moving window PLS [15] and model population analysis [16] have been developed and have produced good results. Not a single absorption band but a region of absorption bands would be reasonable to represent the absorbance of a representative band of pure components and mixture components. Thus, we apply genetic-algorithm-based wavelength selection (GAWLS) method [17] which is a wavelength-regions selection method to IOT. GAWLS is one of the methods used to select combinations of important wavelengths using regions as a unit of measurement. A genetic algorithm (GA) [18] is applied to select wavelength-regions. A GA is an optimization method that is used in biology to model principles of natural evolution. Species having a high level of fitness under certain environmental conditions can prevail in the next generation, and the best species may be reproduced by crossover together with the random mutation of chromosomes in those species that survive. The solution space around superior individuals is searched for preferentially, which leads to the discovery of a solution that is close to the optimum. IOT is performed to calculate a fitness value. This method is named GAWLS-IOT.

In GAWLS-IOT, two actual values of a chromosome represent one region of wavelengths. Fig. 2 shows the basic concept of GAWLS-IOT. Hence, GAWLS-IOT can select important wavelengths using regions as

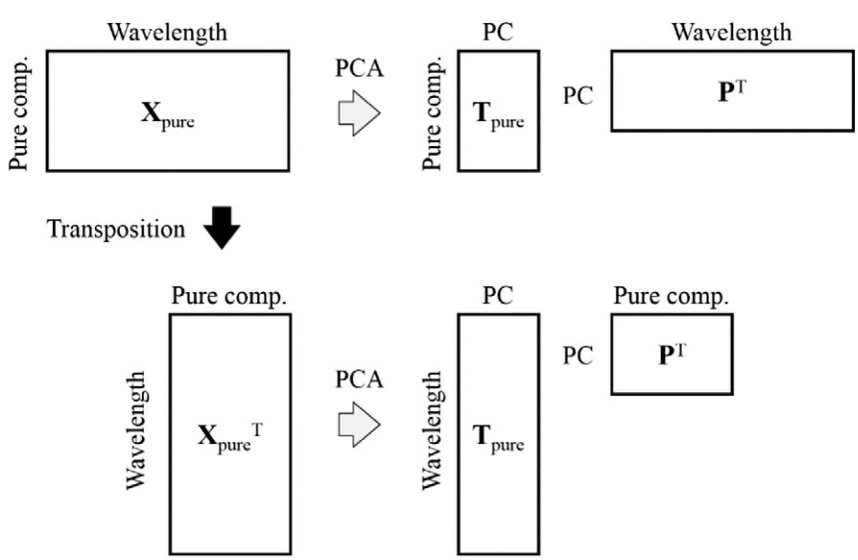

Fig. 1. Basic concept of two ways to apply PCA.